FINDING SIMILAR ENTITIES ACROSS KNOWLEDGE GRAPHS

Sareh Aghaei and Anna Fensel

Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Innsbruck, Austria

ABSTRACT

Finding similar entities among knowledge graphs is an essential research problem for knowledge integration and knowledge graph connection. This paper aims at finding semantically similar entities between two knowledge graphs. This can help end users and search agents more effectively and easily access pertinent information across knowledge graphs. Given a query entity in one knowledge graph, the proposed approach tries to find the most similar entity in another knowledge graph. The main idea is to leverage graph embedding, clustering, regression and sentence embedding. In this approach, RDF2Vec has been employed to generate vector representations of all entities of the second knowledge graph and then the vectors have been clustered based on cosine similarity using K medoids algorithm. Then, an artificial neural network with multilayer perception topology has been used as a regression model to predict the corresponding vector in the second knowledge graph for a given vector from the first knowledge graph. After determining the cluster of the predicated vector, the entities of the detected cluster are ranked through sentence-BERT method and finally the entity with the highest rank is chosen as the most similar one. To evaluate the proposed approach, experiments have been conducted on real-world knowledge graphs. The experimental results demonstrate the effectiveness of the proposed approach.

KEYWORDS

Knowledge Graph, Similar Entity, Graph Embedding, Clustering, Regression, Sentence Embedding.

1. INTRODUCTION

With the rise of knowledge graphs (KGs), interlinking KGs has attracted a lot of attention. A KG is a huge semantic net which integrates various, inconsistent and heterogeneous information resources to represent knowledge about different domains [1]. KGs have proven beneficial for artificial intelligence applications, including question answering, document retrieval, recommendation systems and knowledge reasoning [2, 3]. To interlink KGs, it is crucial to find similar entities across the KGs that have high semantic similarity to each other [3]. Addressing this challenge would allow end users and search agents to find more relevant information across KGs [3]. This can be used in different applications, such as online marketing, search engine optimisation and online services provisioning, for example, in tourism [4].

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 01-11, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110301

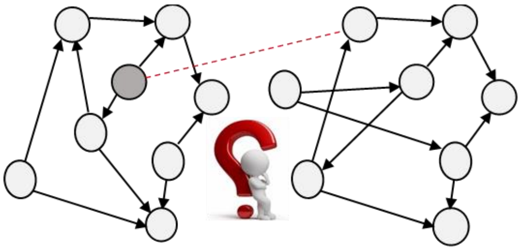

Figure 1. The interlinking problem over the knowledge graphs

This paper delves into the problem of finding similar entities across different KGs. Given a query entity in one KG, this study aims to find the most similar entity in another KG as illustrated in Figure 1. Here, the entity pair may not reference the same real-world entity but have the most similarity to each other. The proposed approach includes four main steps: graph embedding, clustering, regression, sentence embedding as showed in Figure 2.

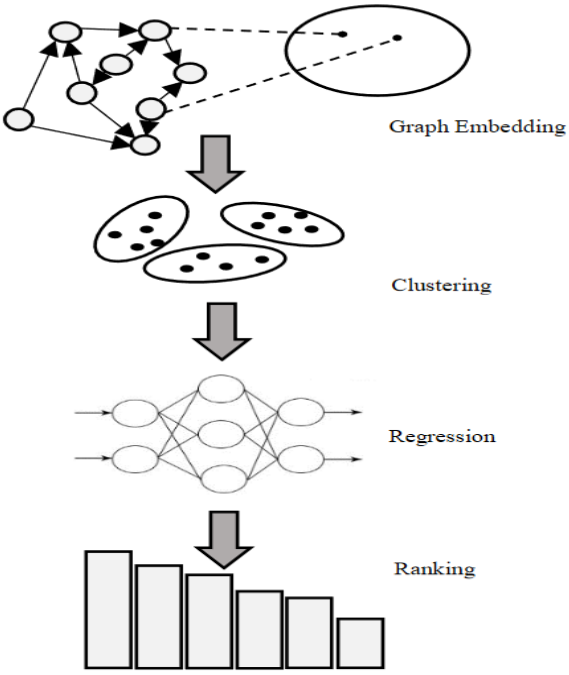

Figure 2. The proposed approach

A graph embedding is used to represent entities of a KG in low dimensional semantic space while preserving the structural as well as the semantic features of the entities. Recently, different graph embedding techniques have been proposed to capture different aspects of graphs. In this paper, RDF2VEC graph embedding technique has been applied to capture the semantic similarity of entities in each RDF KG. RDF2VEC adapts the language modelling approach of word2vec to RDF graph embeddings [5]. RDF2Vec converts the KG to a set of sequences (using graph walks and Weisfeiler-Lehman subtree RDF graph kernels) and then trains a neural network model to learn vector representation of entities. It maps each entity to a low dimensional vector of latent numerical values in which semantically and syntactically closer entities will appear closer in the vector space [5, 6].

Clustering for interlinking large-scale KGs is a fundamental step. Although there are different approaches for clustering large amounts of data, the proposed approach uses K means and K medoids clustering based on two different metrics, Euclidean distance and cosine distance, to group vector representations of the second KG. In these algorithms, vectors are segmented into different groups where each cluster contains at least one vector. No vectors may be placed into more than one cluster. Furthermore, the number of clusters 'K' must be specified prior to initiating the algorithm and also, they allow for interpretability of the cluster centres. K-means targets to minimize the total squared error from a central position in each cluster namely centroid. Whereas K medoids aims to minimize the sum of dissimilarities between vectors labelled to be in a cluster and one of the vectors considered as the representative of that cluster called medoid [7, 8].

Various methods, including a variety of regression techniques and artificial neural networks, can be applied to develop a forecasting model. The present approach has employed artificial neural network and multivariate multiple linear regression techniques to predict the vector representation for a given embedding from the first KG. The neural network technique has become an increasingly popular modelling tool for forecasting. Multilayer perceptron (MLP) with back-propagation learning rule is adopted to predict the embeddings of the second KG according to the embeddings of the first KG. Furthermore, the multivariate multiple linear regression model is beneficial in discovering the association between various independent and dependent variables. It attempts to model the correlation between involving variables and response variables depending on linear equation into the observed data [9].

The entity description and other textual values of properties in KG usually carry conceptual semantic information [10]. Based on the entity description, the Sentence-BERT technique is adopted to compute the textual similarity. Sentence-BERT (S-BERT) is a modification of the pretrained BERT network that employs siamese and triplet networks in order to derive semantically meaningful sentence embeddings [11]. The derived sentence embeddings of the entities of the chosen cluster are compared with the sentence embedding of the given entity and ranked based on cosine-similarity. Finally, the entity ranked first is selected.

The approach of this paper can take advantage of value-oriented and record-oriented [12] techniques. According to [12], value-oriented techniques compute the similarity between entities on the attribute level and record-oriented techniques contain solutions based on learning, rules, contexts. Furthermore, it works independent of mapping schema and benefits the structure of KGs.

The remainder of this paper is structured as follows. The next section presents some related studies. In section 3, the proposed approach is presented. Section 4 demonstrates the results obtained and evaluation. Finally, concluding remarks and an outlook on future work are in Section 5.

2. RELATED WORKS

The task of interlinking KGs aims to find entities in two KGs that have semantic relations. The different KGs are constructed independently from each other, so they contain complementary entities. While numerous studies exist regarding entity alignment (also named entity resolution, duplicate detection, record linkage, or entity resolution) with the goal of finding entities from different KGs that refer to the same real-world identity [9], there is a lack of approaches to find entities with the most similarity so that those entities may not be the same entity pairs.

SILK [13], LIMES [14] and Dude [15]are examples of traditional approaches which have leveraged different similarity metrics including string similarity, numeric similarity, date similarity, word relation and fuzzy string similarity. These approaches usually have an ability to build more complex similarity metrics through combining the similarity metrics for increasing their functionality and performance.

In [3], a classification-based approach has been provided to address the entity alignment problem between source and target KGs. Using source/target entity pairs, a classifier is trained and the probability of predicting an alignment is adopted for candidate ranking. RDF2Vec graph embedding technique has been used to the embeddings of the source and target entities, then the embedding of the given entity in the source KG and the candidate entity in the target KG are concatenated into one feature vector and fed into a multi-layer perception. Finally, it sorts the candidates by the match probability for evaluation.

MtransE [16] which is a multi-lingual KG embedding model has consisted of two component models, called knowledge model and alignment model, to learn the multilingual KG structure. The knowledge model encodes entities by adopting TransE [17]. On top of that, the alignment model employs three different techniques to learn cross-lingual alignment for entities and relations, namely distance-based axis calibration, translation vectors, and linear transformations. Comparisons across the used techniques show that the linear-transformation-technique based on different loss functions.

A KG alignment network, namely AliNet [18] has been proposed to reduce the non-isomorphism of neighbourhood structures in an end-to-end manner. Since the schema heterogeneity ensures dissimilarity across counterpart entities, AliNet introduces distant neighbours to expand the overlap between their neighbourhood structures using an attention mechanism. The neighbourhood information within multiple hops are captured through the applied gating mechanism in each layer.

For cross-lingual entity alignment, a joint attribute-preserving embedding model has been introduced to jointly embed the structures of two knowledge bases into a unified vector space and then refine it through leveraging attribute correlations in the knowledge bases. This model has utilized the structure embedding and attribute embedding in order to represent the relationship structures and attribute correlations of knowledge bases and learn approximate embeddings for latent aligned entities [19].

REA [20] has proposed a framework for robust entity alignment over KGs. The framework consists of two components: noise detection and noise-aware entity alignment. In order to encode the information of KGs, it leverages a graph neural network-based encoder. The noise-aware entity alignment component targets to diminish the distance between two entities in a labelled entity pair to avoid the noise based on the encoder. The idea of the noise detection component is to generate noisy data and have an ability to differentiate between the generated noisy data and real data following the adversarial training principle. However, REA cannot distinguish a few real entity pairs with real pairs in some cases.

3. APPROACH

Problem Definition – A Resource Description Framework(RDF)KG can be denoted as $G = (E, R, T)$, where E is the set of entities, R is the set of relations, and T is the set of triples. A KG triple ($\mathrm{e_h}, r, \mathrm{e_t}$ indicates the head entity $\mathrm{e_h}$ is linked to the tail entity $\mathrm{e_t}$ by the relation r. Let $G_1 = (E_1, R_1, T_1)$ and $G_2 = (E_2, R_2, T_2)$ be the first and second KG, respectively. The task is to find the entity $e_2 \in E_2$ which has the most semantic similarity to the given entity $\mathrm{e}_1 \in \mathrm{E}_1$ from the first KG, thus $\forall e_1 \in E_1 : \exists e_2 \in E_2 that\ e_2 \approx e_1$.

Methodology - The proposed approach includes four main steps: graph embedding, clustering, regression, sentence embedding. In the first step, RDF2Vec [6] algorithm has been used to generate RDF graph embeddings. The generated vector representations of the second KG are clustered in the next step. Then, a regression model is trained according to the vector representations of the same entities between the first and second KGs. For each given entity of the first KG, the correspondent vector from the second KG is predicated and its cluster is determined. In the final step, the sentence embedding is utilized based on the value of description property in the predicated cluster by BERT and the generated vectors are ranked based on cosine-similarity with the sentence vector of the source entity. The target entity with top rank is the entity with more similarity.

Below, the approach steps including graph embedding, clustering, regression and rankling are described in detail.

3.1. Graph Embedding

RDF2Vec, which is a technique to embed RDF graphs for learning latent numerical representations of entities in RDF graphs, has been inspired by the word2vec approach. The Word2vec is a particularly computationally-efficient two-layer neural language model to generate word embeddings from raw text [6, 21]. The Word2vec takes a set of sentences as input, and trains a two-layer neural network using one of the two algorithms, the continuous bag of words model (CBOW) and the skip-gram model (SG). The CBOW predicts a target word from its context within a given window and the SG predicts the context words given a word. The RDF2Vec first converts the RDF graphs in a set of sequences using two techniques, Weisfeiler-Lehman Subtree RDF Graph Kernels and graph walks, which are then used as input for the word2vec algorithm to train the neural language model [21]. When the training is done, all entities are projected into a lower-dimensional feature space, and semantically similar entities are closer in the vector space than dissimilar ones. For more details the readers are referred to [6, 21]. In the proposed approach, RDF2Vec is used to generate embeddings for all entities of the second KG, the entity pairs which have the same relation between the first and second KGs and each given entity from the first KG.

3.2. Clustering

Clustering is of key importance for interlinking entities from multiple KG. To achieve high efficiency for large KGs, interlinking solutions have to avoid comparing each entity to all other entities. This can be gained by so-called blocking strategies where only entities within the same cluster (block) need to be compared with each other [22]. Clustering algorithms typically try to cluster entities such that the similarity between entities within a cluster is maximized while the similarity between entities of different clusters is minimized [22].

In proposed approach, the K medoids algorithm has been adopted to cluster vector representations of the second KG. This algorithm is relatively simple to implement and scales to large KGs. Moreover, the medoids of the K clusters can be used to determine the relevant cluster of new vectors. The cosine similarity has been chosen to group together all close vectors of the second KG.

3.3. Regression

The objective of the regression prediction model is to find the transitions between the vector spaces of the first and second KGs [16]. Since the embeddings of the KGs are learned separately, it is essential to learn correspondences between two semantic spaces. One feasible solution to the dilemma is to estimate regression relationships between the entities of the first KG and the entities of the second KG based on existing similar entities.

In this step, the following regression prediction models have been applied:

Multi-layer perceptron (MLP) network - One of the most popular artificial neural networks which can be used to find associations between two sets of variables, is the feed-forward multi-layer network, which uses a back-propagation learning algorithm. It consists of one or more hidden layer(s), containing computational nodes named neurons/perceptrons which intervene between input and output of the network, and can improve the accuracy of the network [23].

Multivariate multiple linear regression (MMLR) - The multivariate multiple linear regression is a statistical method that allows to predict of several dependent variables from a set of independent variables and its purpose is finding the best fitting line which is called regression function [24, 25].

In the proposed approach, a MLP network and also a MMLR is used to predict the embedding of similar entity in the second KG based on the embedding of the given entity of the first KG. The entity pairs which have same relation between the first and second KGs have been considered as training data to train the regression models.

3.4. Ranking

Sentence-BERT (SBERT) can be considered a modification of the pretrained BERT network which generates a fixed sized sentence embedding by adding a pooling operation to the output of BERT / RoBERTa. In order to fine-tune BERT / RoBERTa, SBERT uses siamese and triplet networks to update the weights such that the generated sentence embeddings are semantically meaningful and can be compared with cosine-similarity [11].

In the proposed approach, the textual values of entity properties (e.g. description) are adopted to input sentences for SBERT. The sentence embeddings of the determined cluster in the previous step are compared with the sentence embedding of the given entity from the first KG and then ranked based on cosine similarity. Finally, the highest ranked entity is chosen as the entity which has the most semantic similarity with the given one.

4. EVALUATION

In order to evaluate the approach presented in this paper, DBPedia [26] and SalzburgerLand [27] KGs have been used as the first and second KGs. The SalzburgerLand KG is a KG describing touristic entities of the region of Salzburg, Austria, and among others it includes 21496 triples and 571 entities which reference DBPedia KG, which is a KG representing Wikipedia. The evaluation code has been written in Python and is publicly available athttps://github.com/sareaghaei/interlinking.

For RDF2Vec graph embedding, the depth of graph walks and the limit number of walks per entity are 8 and 20, respectively. The outcome of this step is a 100-dimensional vector for each entity.

K means and K medoids algorithms have been employed to group together all close vectors of SalzburgerLand KG based on the Euclidean distance and the standard cosine similarity, respectively. In practice, for obtaining the best clustering quality, the optimal value of K is determined by experiments (K = 2). Not only does K medoids clustering has higher score in terms of silhouette coefficient, but also leads to better result in the next steps. The figure for silhouette coefficient is 0.60 and 0.39 in K medoids and K means, respectively. Figure 3 illustrates the sets of clustering. Also, the centroids and medoids have been shown in green colour and bigger size.

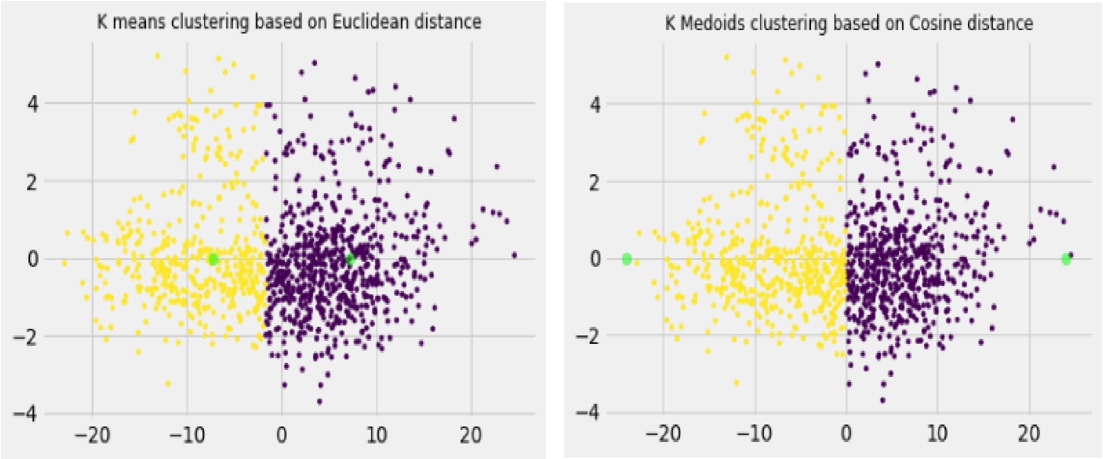

Figure 3. The clusters of K means and K medoids algorithms

Note that principal component analysis (PCA) [28] has been used to automatically perform dimensionality reduction over the embeddings before visualizing the clusters in Figure 3. The Scikit-Learn library, which has implemented the PCA technique, applies the full singular value decomposition (SVD) or a randomized truncated SVD depending on the shape of the input data and the number of components to extract [29]. Here, PCA has been used to reduce dimensions from 100 to 2.

A multi-layer perception (MLP) with 1 hidden layer which has size 50 using the ReLU activation function, followed by a fully-connected layer and ReLU to output the final prediction has been applied as the regression predication model. The model is trained using the Adam optimizer. Moreover, a multivariate multiple linear regression model has been trained in which the DBPedia KG and the SalzburgerLand KG embeddings are considered as independent and dependent variables, respectively. In order to provide a more complete and effective evaluation of the regression models, the cross-validation has been performed using K-fold algorithm with 5-folds.

The smaller the difference between the predicted vectors and the real vectors, the higher the prediction accuracy that the models provide. Thus, mean absolute error (MAE), mean square error (MSE) and root mean square error (RMSE) have been applied to measure the performance of the models. MAE represents the average of the absolute difference between the actual and predicted values, MSE is defined as average of the square of the difference between actual and predicted values and RMSE is the square root of mean squared error which computes the standard deviation of residuals. Mathematical formulas to calculate these metrics can be written as following where $y^{\wedge}$ is the predicated value of y:

$$MAE = \frac{1}{n}\sum_{i=1}^{i=n} |y_i - y_i^{\wedge}|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{i=n} (y_i - y_i^{\wedge})^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{i=n} (y_i - y_i^{\wedge})^2}$$

Overall, the MLP network outperforms the MMLR model for predicting the embeddings of SalzburgerLand KG, the figures for MAE and MSE in the MLP network are 0.866 and 1.005, respectively, whereas those of the MMLR model are 0.966 and 1.348, the evaluation of results is shown in Figure 4.

Sentence-Transformers which is a Python framework has been used to compute sentence / text embeddings [11]. It is based on PyTorch and Transformers and the produced sentence embeddings are 150-dimensional vectors which have been compared with cosine-similarity in order to be ranked.

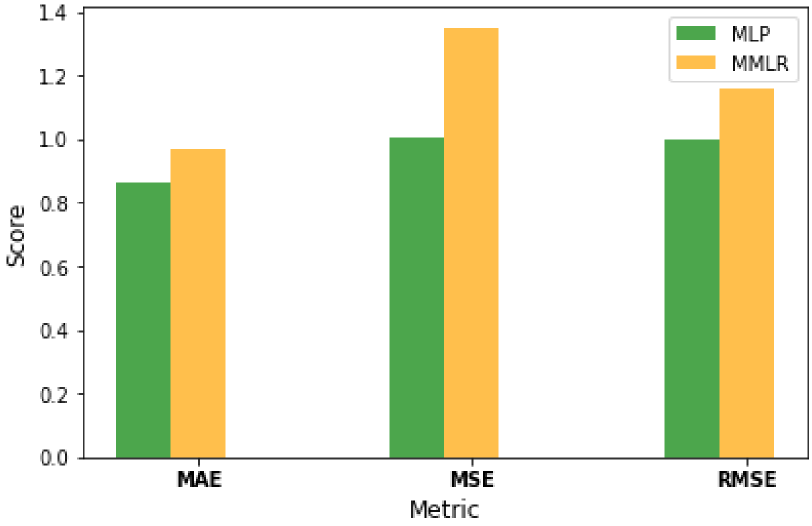

Figure 4. The evaluation of prediction errors

5. CONCLUSION

This paper proposes an approach to interlink KGs. In order to find the most similar entity from a KG (second KG) with a given entity from another KG (first KG), the proposed approach includes four steps: graph embedding, clustering, regression and ranking. RDF2Vec technique is used to generate vector representations and then K means/K medoids algorithms are adopted for clustering of the embeddings of the second KG. To learn associations between distinct semantic spaces (one from each KG), multi-layer perceptron networks and multivariate multiple linear regressions are trained and used to predict the embedding from the second KG based on the embedding of the given entity from the first one. By comparing the predicted vector with the centroid-medoid of the clusters, the correspondent cluster is determined and its entities are ranked based on cosine similarity between their sentence embedding and the sentence embedding of the given entity. SBERT is used to compute sentence embeddings of the entities over their textual values of the properties. The experimental results show that the proposed approach as one of the state-of-art interlinking approaches can achieve high accuracy. However, in the proposed approach, the regression model requires training based on the entity pairs between two KGs, it can definitely be considered a drawback due to lack of the pairs in some cases. For future work, aside from experimenting with other embedding learning techniques for KGs, learning associations on KGs with better accuracy and experiments on different KGs are planned.

ACKNOWLEDGEMENTS

This work has been partially funded by the project WordLiftNG within the Eureka, Eurostars Programme (grant agreement number 877857 with the Austrian Research Promotion Agency (FFG)).

REFERENCES

[1] Dieter Fensel, Umutcan Simsek, Kevin Angele, Elwin Huaman, Elias Kärle,Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler(2020) “Introduction: What Is a Knowledge Graph?”, Springer International Publishing, pp. 1–10.

[2] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu (2018) “Bootstrapping entity alignment with knowledge graph embedding”, International Joint Conferences on Artificial Intelligence, pp. 4396-4402.

[3] Michael Azmy, Peng Shi, Jimmy Lin, and Ihab F. Ilyas (2019) “Matching entities across different knowledge graphs with graph embeddings”, CoRR, abs/1903.06607.

[4] Anna Fensel, Zaenal Akbar, Elias Kärle, Christoph Blank, Patrick Pixner, and Andreas Gruber (2020) “Knowledge Graphs for Online Marketing and Sales of Touristic Services”, Information, 11(5), 253.

[5] RemziCelebi, HuseyinUyar, Erkan Yasar, OzgurGumus, OguzDikenelli, and Michel Dumontier (2019) “Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings”, BMC Bioinformatics.

[6] PetarRistoski, and Heiko Paulheim (2016) “RDF2Vec: RDF Graph Embeddings for Data Mining.” International Semantic Web Conference.

[7] Tagaram Soni Madhulatha (2011) “Comparison between K-Means and K-Medoids Clustering Algorithms”, Advances in Computing and Information Technology, pp. 472-481.

[8] Preeti Arora, DeepaliVirmani, and Shipra Varshney (2016) “Analysis of K-Means and K-Medoids Algorithm For Big Data”, Procedia Computer Science, Vol. 78, pp. 507-512.

[9] Elwin Huaman, Elias Kärle, and Dieter Fensel (2020) “Duplication Detection in Knowledge Graphs: Literature and Tools”, arXiv:2004.08257.

[10] Ying Shen, Kaiqi Yuan, Jingchao Dai, Buzhou Tang, Min Yang, and Kai Lei (2019) “KGDDS: A System for Drug-Drug Similarity Measure in Therapeutic Substitution based on Knowledge Graph Curation”, Journal of medical systems 43, 92.

[11] Nils Reimers, and Iryna Gurevych (2019) "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982-3992.

[12] Silvana Castano, Alfio Ferrara, Stefano Montanelli, and Gaia Varese (2011) "Ontology and instance matching. In Knowledge-Driven Multimedia Information Extraction and Ontology Evolution", Springer, pp. 167-195.

[13] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov (2009) "Discovering and Maintaining Links on the Web of Data", In Proceedings of The International Semantic Web Conference (ISWC) ISWC, pp. 650-665.

[14] Axel-CyrilleNgongaNgomo, and Sören Auer (2011) "LIMES - A time-efficient approach for large-scale link discovery on the web of data", In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI2011), pp. 2312-2317.

[15] Lars Marius Garshol, and Axel Borge (2013) "Hafslundsesam - an archive on semantics", In Proceedings of the 10th Extending Semantic Web Conference (ESWC2013), vol. 7882, pp. 578-592.

[16] Muhao Chen, Yingtao Tian, MohanYang, and Carlo Zaniolo (2016) "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment", arXivpreprint arXiv:1611.03954.

[17] Antoine Bordes,NicolasUsunier, Alberto Garcia-Duran, Jason Weston, and OksanaYakhnenko (2013) "Translating embeddings for modelling multi-relational data", In Proceedings of the 26th International Conference on Neural Information Processing Systems, pp. 2787-2795.

[18] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu (2019) "Knowledge graph alignment network with gated multi-hop neighborhood aggregation", arXiv:1911.08936.

[19] Zequn Sun, Wei Hu, and Chengkai Li (2017) "Cross-lingual entity alignment via joint attribute-preserving embedding", In Proceedings of The International Semantic Web Conference (ISWC) ISWC, vol. 10587, pp. 628-644.

[20] Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang (2020) "Rea: Robust cross-lingual entity alignment between knowledge graphs", In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery& Data Mining, pp. 2175-2184.

[21] PetarRistoski, Jessica Rosati, Tommaso Di Noia, Renato De Leone, and Heiko Paulheim (2018) "RDF2Vec: RDF Graph Embeddings and Their Applications", Semantic Web, Vol. 10, No. 4, pp. 721-752.

[22] Ali Saeedi, Markus Nentwig, Eric Peukert, and Erhard Rahm (2018) "Scalable matching and clusteringof entities with famer", Complex Systems Informatics and Modelling Quarterly, pp. 61-83.

[23] M.E. Hamzehie, S Mazinani, F. Davardoost, A. Mokhtare, H. Najibi, BVdBruggen, and S. Darvishmanesh (2014) "Developing a feed forward multilayer neural network model for prediction of CO2 solubility in blended aqueous amine solutions", Journal of Natural Gas Science and Engineering, pp. 19-25.

[24] Lianpeng Li, Jian Dong, Decheng Zuo, and Jin Wu (2019) "SLA-aware and energy-efficient VM consolidation in cloud data centers using robust linear regression prediction model", IEEE Access 7, pp. 9490-9500.

[25] Yanming Li, Bin Nan, and Ji Zhu (2015) "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure", Biometrics, 71, pp. 354-363.

[26] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer (2013) "A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, pp. 167-195.

[27] "Welcome to SalzburgerLand Data Hub", Accessed on: Oct. 18, 2020. [Online]. Available: http://data.salzburgerland.com/dataset/salzburgerland-en.

[28] IanT. Jolliffe, and Jorge Cadima (2002) "Principal Component Analysis", Wiley Online Library.

[29] "Principal component analysis (PCA)", Accessed on: Oct. 5, 2020. [Online]. Available: https://scikit-learn.org/stable/modules/decomposition.html#pca.

AUTHORS

Sareh Aghaei received the master's degree in computer engineering from the University of Isfahan, Iran and is currently a PhD student at the University of Innsbruck, Austria. Her research areas include semantic web, knowledge graphs and question answering systems.

Anna Fensel is Associate Professor at the University of Innsbruck, Austria. Earlier she worked as a Senior Researcher at FTW – Telecommunications Research Centre Vienna, Austria, and a Research Fellow at the University of Surrey, UK. Anna has earned both her habilitation and her doctoral degree in Computer Science at the University of Innsbruck, and she has a university degree in Mathematics and Computer Science degree from Novosibirsk State University, Russia.